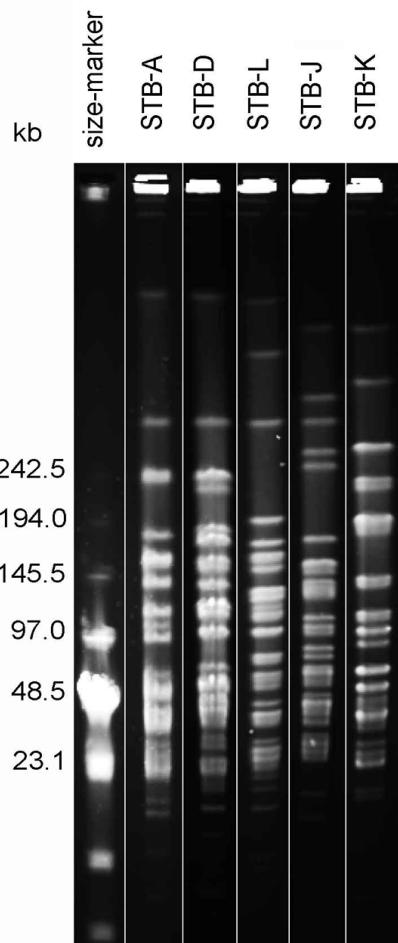


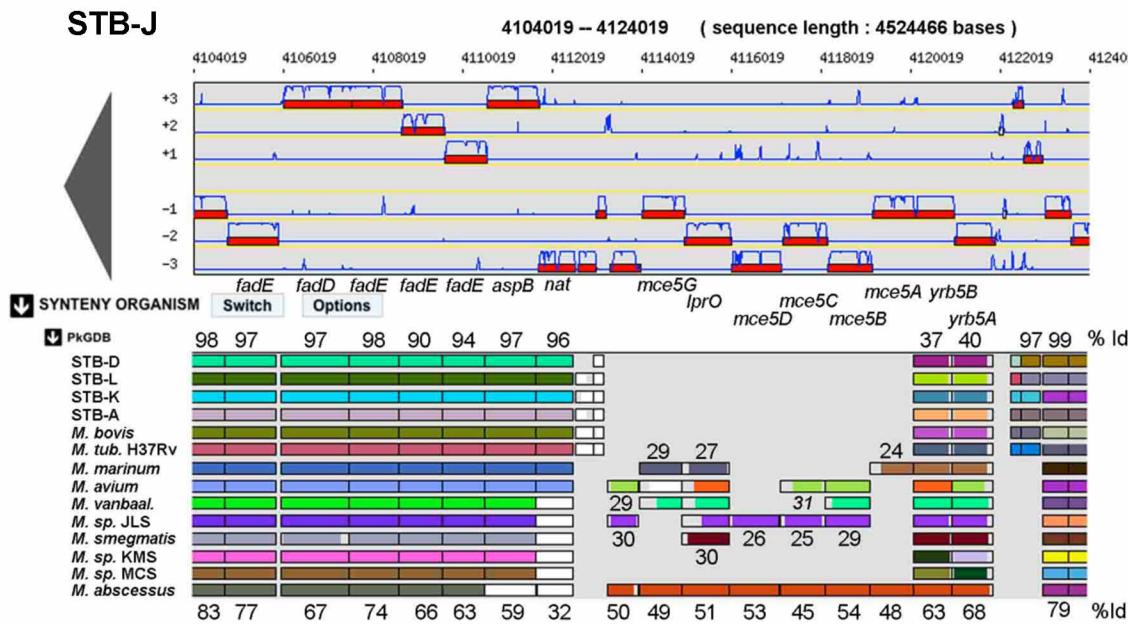
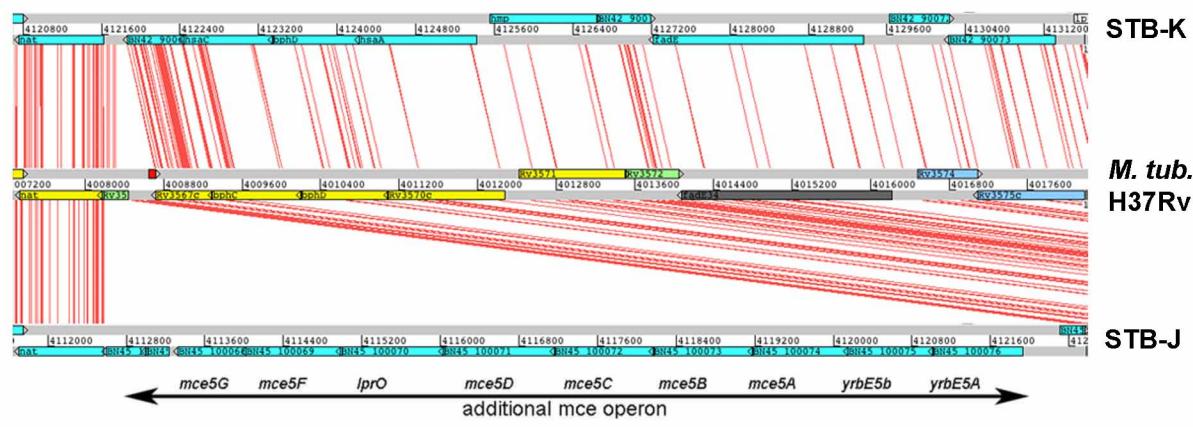
## **Supplementary Information:**

### **Genome analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of the etiologic agent of tuberculosis**

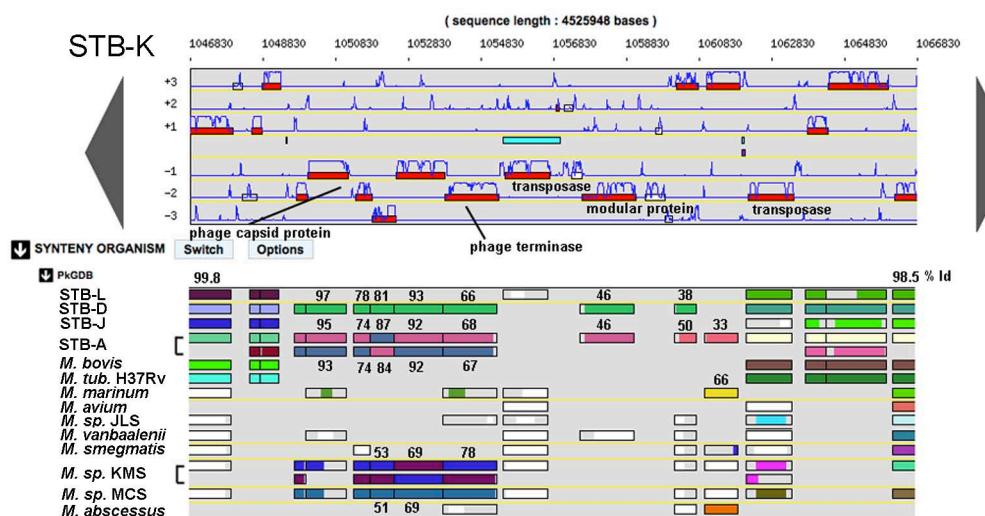
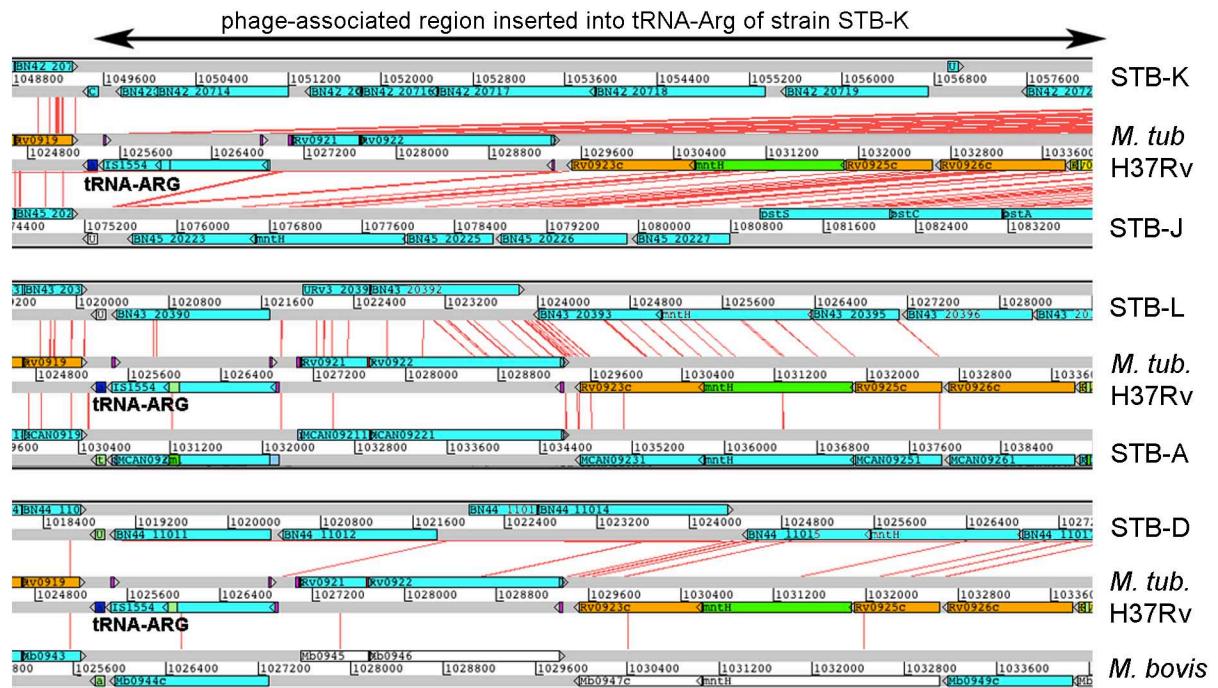
Philip Supply, Michael Marceau, Sophie Mangenot, David Roche, Carine Rouanet, Varun Khanna, Laleh Majlessi, Alexis Criscuolo, Julien Tap, Alexandre Pawlik, Laurence Fiette, Mickael Orgeur, Michel Fabre, Cécile Parmentier, Wafa Frigui, Roxane Simeone, Eva C. Boritsch, Anne-Sophie Debré, Eve Willery, Danielle Walker, Michael A. Quail, Laurence Ma, Christiane Bouchier, Grégory Salvignol, Fadel Sayes, Alessandro Cascioferro, Torsten Seemann, Valérie Barbe, Camille Locht, Maria-Cristina Gutierrez, Claude Leclerc, Stephen Bentley, Timothy P. Stinear, Sylvain Brisse, Claudine Médigue, Julian Parkhill, Stéphane Cruveiller & Roland Brosch.



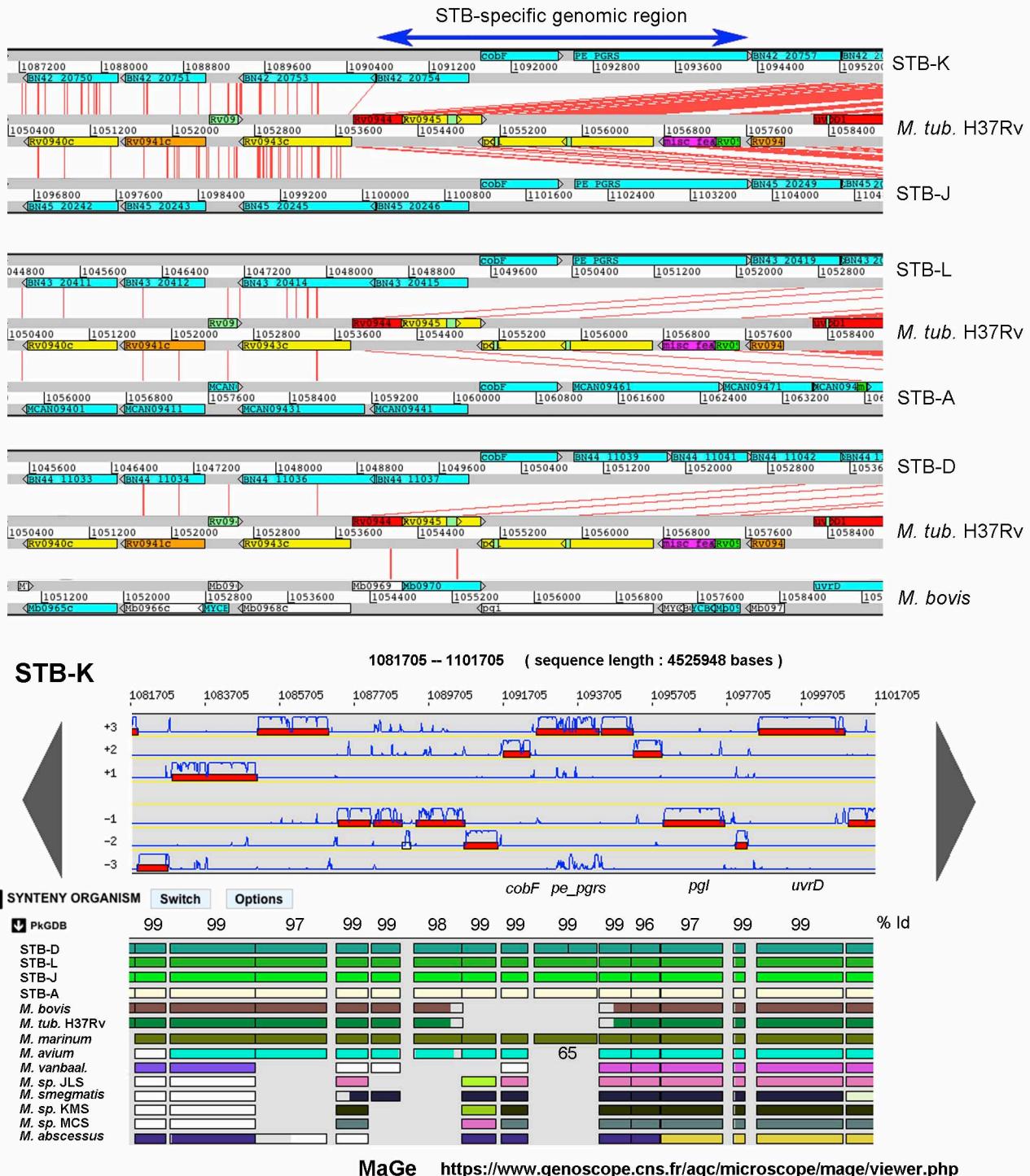
**Supplementary Fig. 1.** Large genomic diversity of smooth tubercle bacilli (STB).  
Asel restriction fragments separated by pulsed-field gel electrophoresis (PFGE) of selected STB strains.



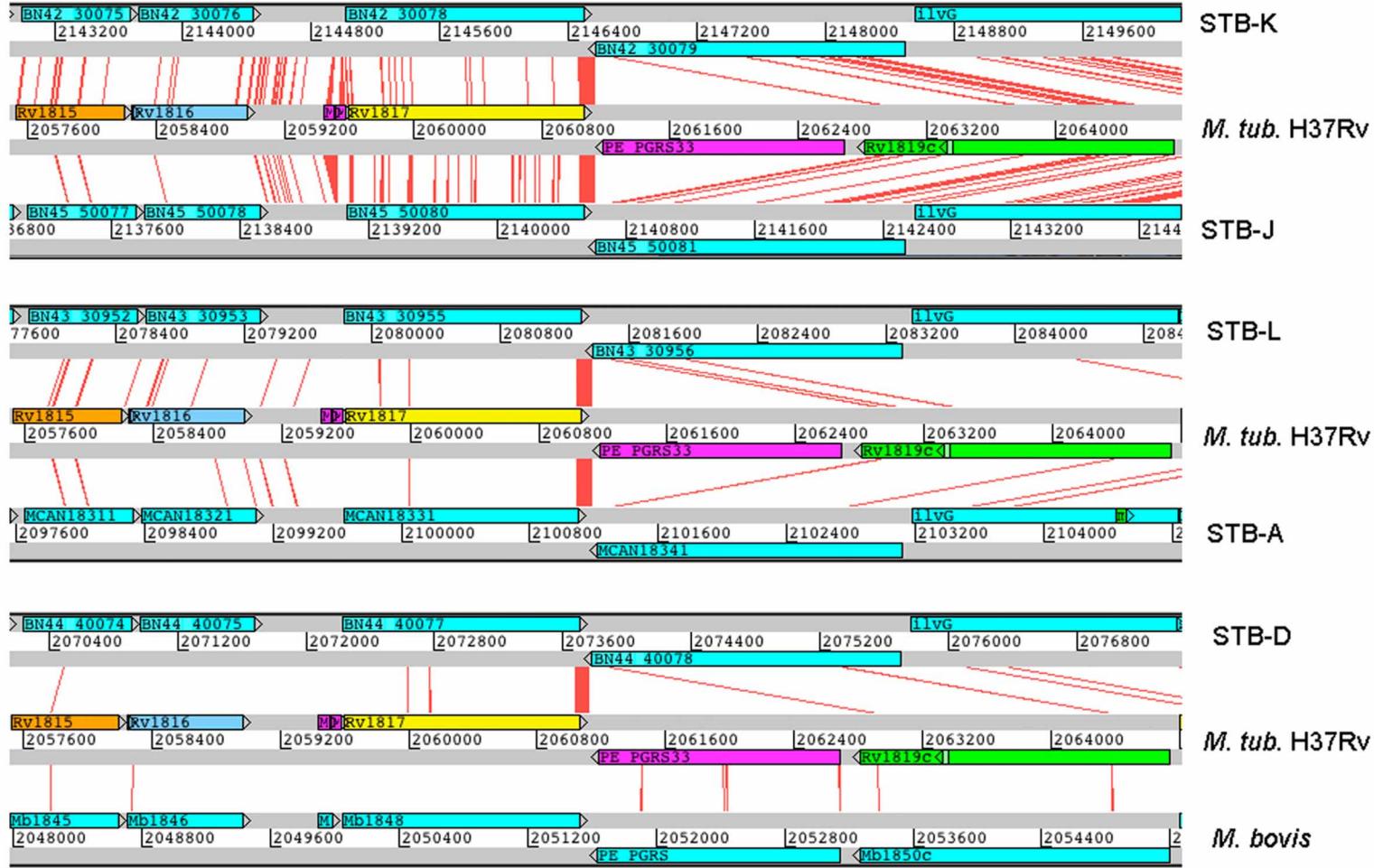
**Supplementary Fig. 2.** Extended pan-genome of tubercle bacilli. (a) Genomic region encoding a novel Mce (mycobacterial cell entry) protein cluster that is present only in STB strain J, as depicted by the Artemis Comparison Tool (ACT, upper panel) and the Magnifying Genome software (MaGe, lower panel). The red lines in the upper panel indicate SNPs identified between pairwise compared genomes. *M. tub.* = *M. tuberculosis*, *M. vanbaal.* = *M. vanbaalenii*.



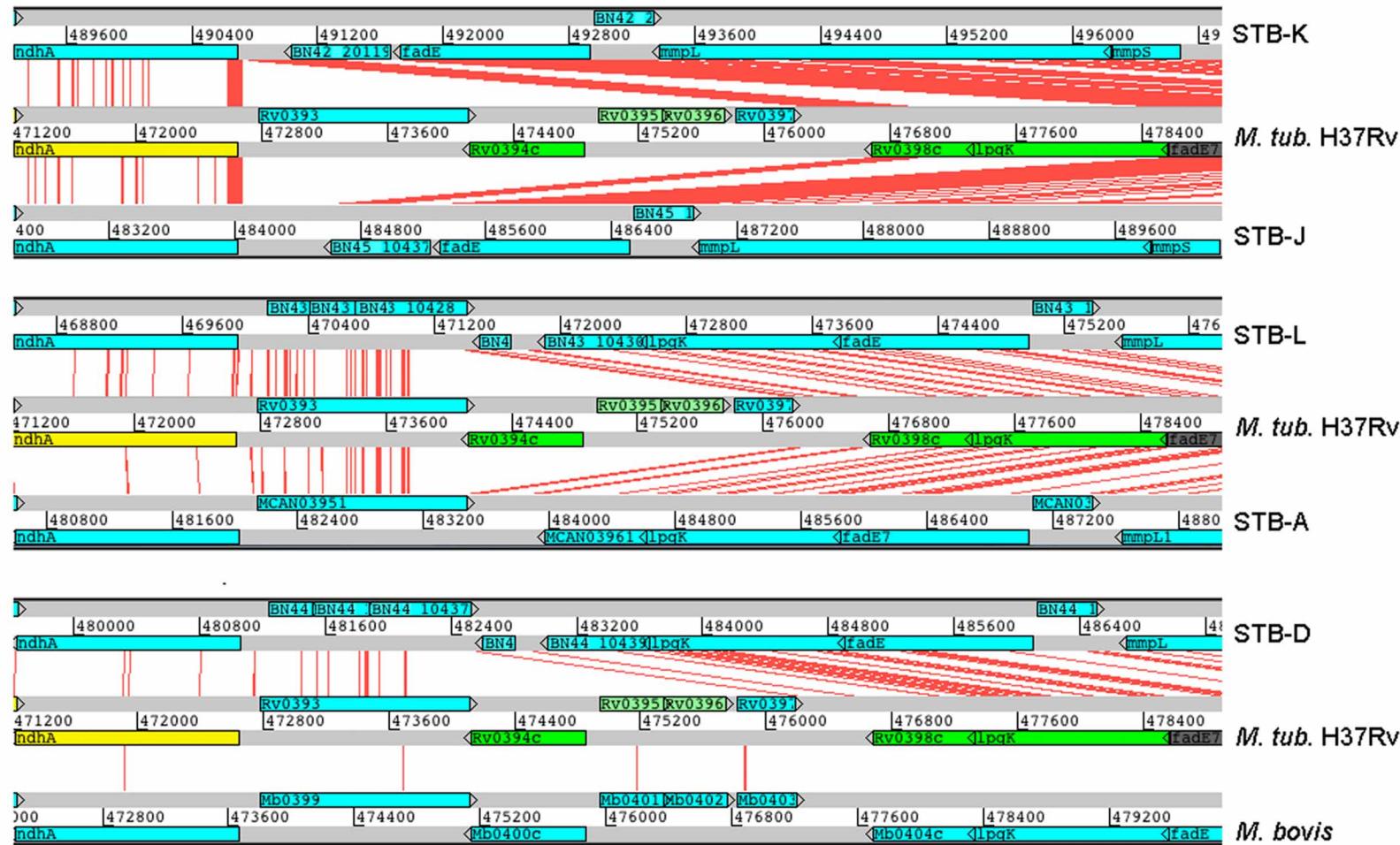
**Supplementary Fig. 2.** Extended pan-genome of tubercle bacilli. **(b)** Example of the presence of a phage-associated gene cluster in STB strain K that is not found in MTBC. Note that STB strains D and A also contain similar gene clusters inserted into another genomic locus as indicated by the color change of the CDS, indicating interruption of synteny (lower panel). Of note, proteins with substantial similarity to these phage-associated proteins are also present in the distantly related, environmental *Mycobacterium* sp. KMS and MCS strains (lower panel).



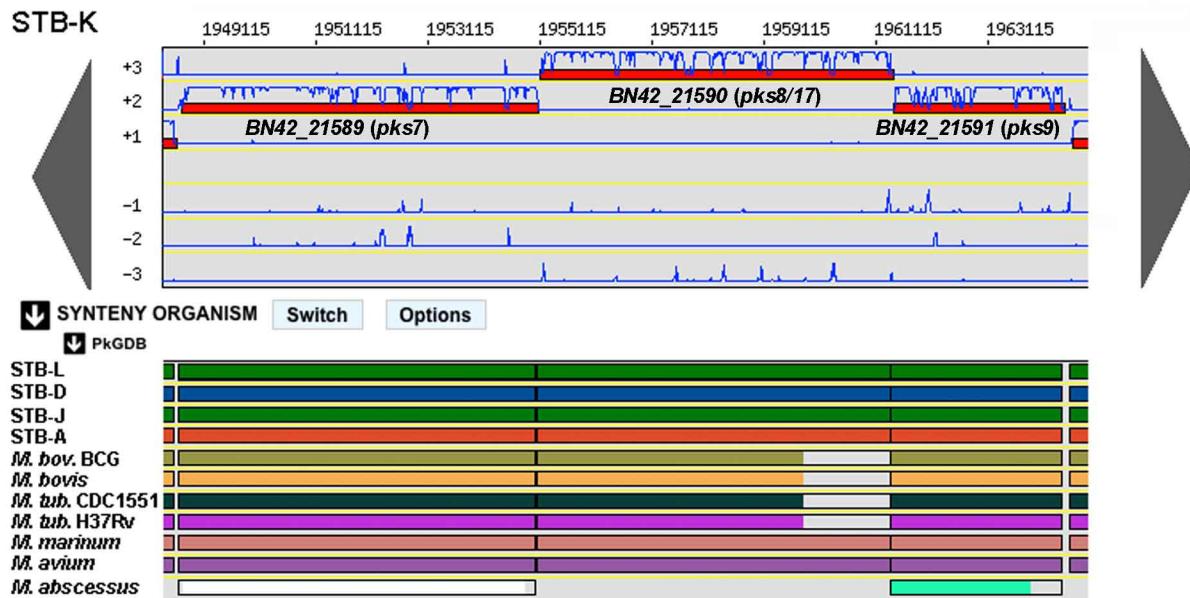
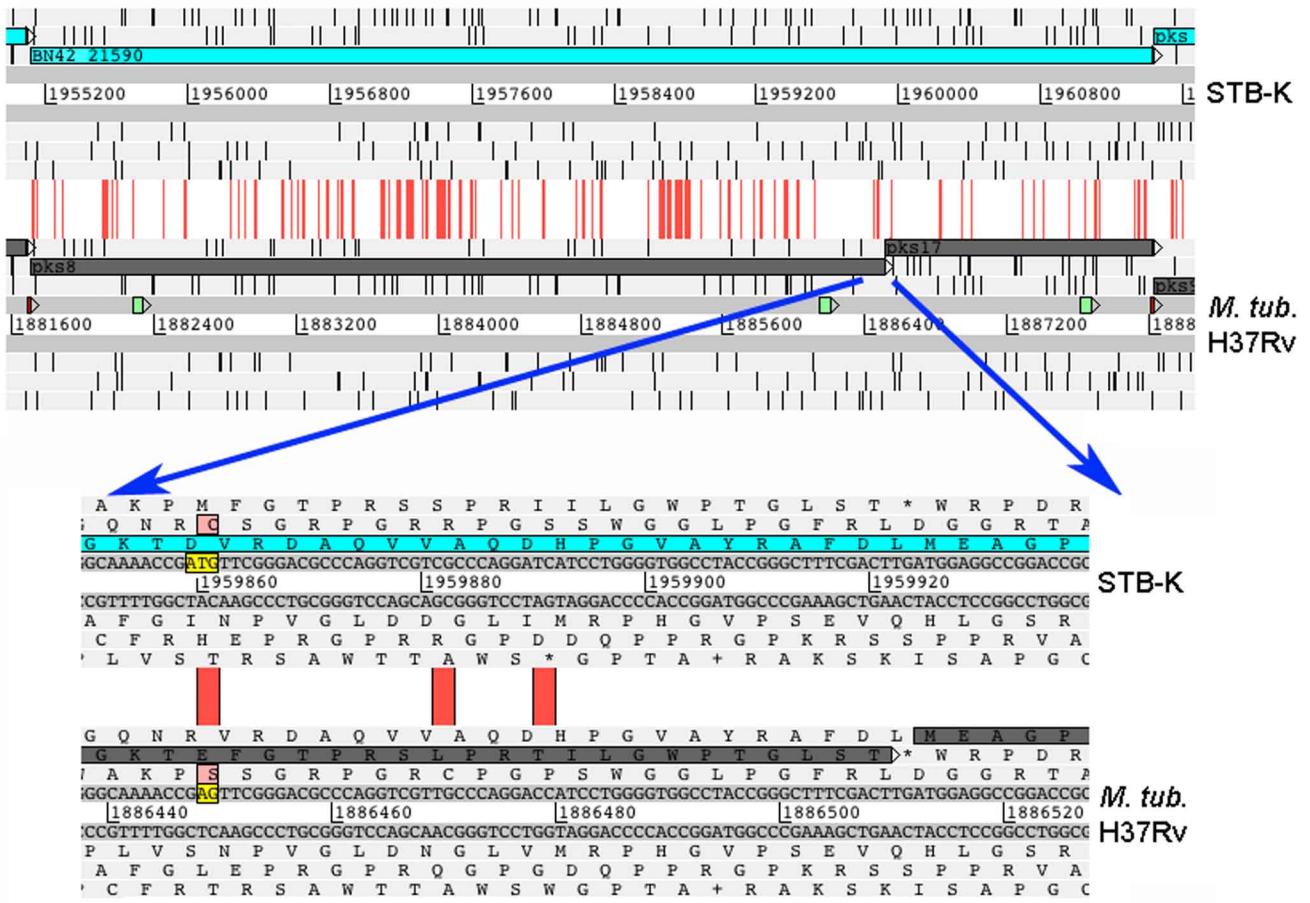
**Supplementary Fig. 2.** Extended pan-genome of tubercle bacilli. (c) Representation of a genomic locus that was found in all STB strains tested, but which was absent from all MTBC strains. The region encodes 3 proteins including a conserved hypothetical protein, a CobF homologue and a specific PE\_PGRS protein. As similar proteins are also present in *M. marinum*, it seems likely that the presence of these three genes corresponds to the ancestral form, followed by deletion of the genes in the lineage of MTBC strains after its divergence from lineages of smooth tubercle bacilli.



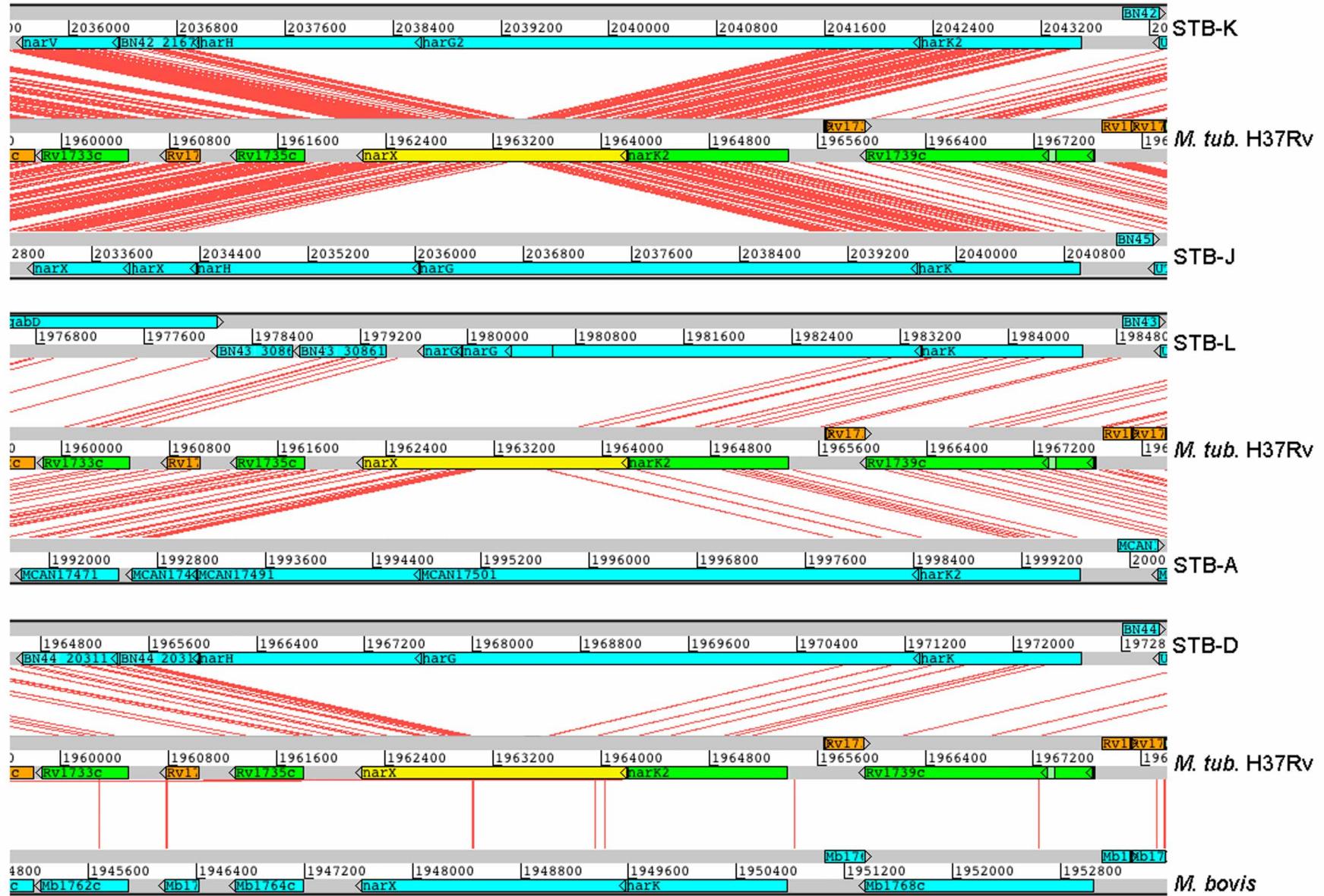
**Supplementary Fig. 3.** *Mycobacterium tuberculosis* complex (MTBC) specific genes. (a) Example of a gene that is specifically present in the MTBC and absent from smooth tubercle bacilli (STB). Comparison by ACT clearly shows that compared to STB genomes, MTBC members harbor the gene encoding protein PE\_PGRS33 at the same genomic location next to gene *rv1817*. Note that in the orthologous region of *M. marinum* a completely different PE\_PGRS encoding gene is present (not shown). From this genomic organization it can be deduced that the gene for PE\_PGRS33 (*rv1818c*), which has repeatedly been described to enhance cellular toxicity during infection, has specifically been inserted into this genomic region of MTBC. *M. tub.* = *M. tuberculosis*



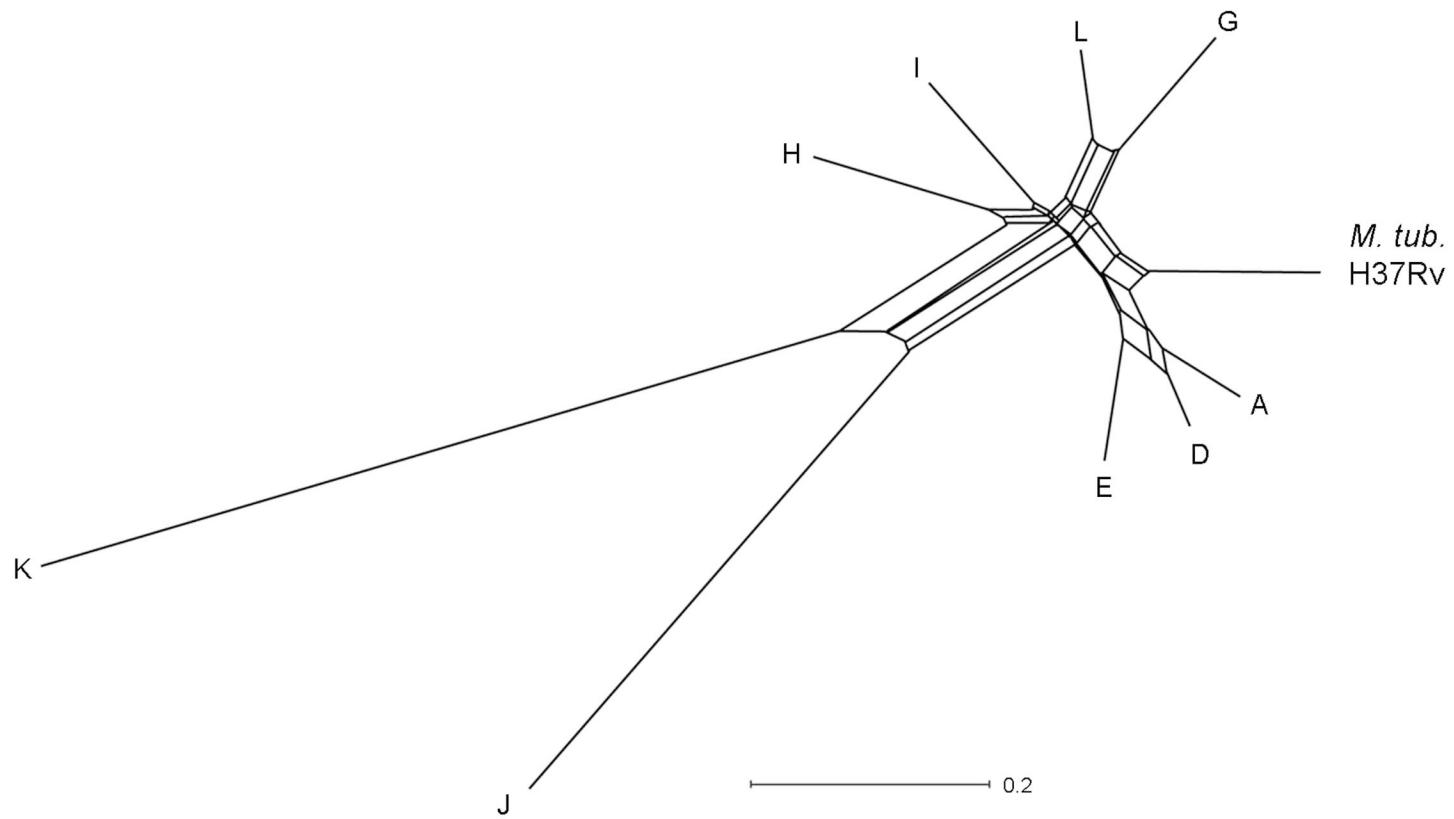
**Supplementary Fig. 3.** *Mycobacterium tuberculosis* complex (MTBC) specific genes. (b) Example of other genes specifically present in the MTBC and absent from STB strains as depicted by ACT comparison. Best BLAST hits (outside the MTBC) for Rv0394 were obtained with hypothetical protein of *Gordonia otitidis* (35% amino acid sequence identity) or *Saccharomonospora marina* (28%), while Rv0395/Rv0396 showed best hits with *Mycobacterium colombiense* (71% and 69% respectively), suggesting horizontal gene transfer from distant donor species into MTBC after separation of the MTBC from STB lineages.



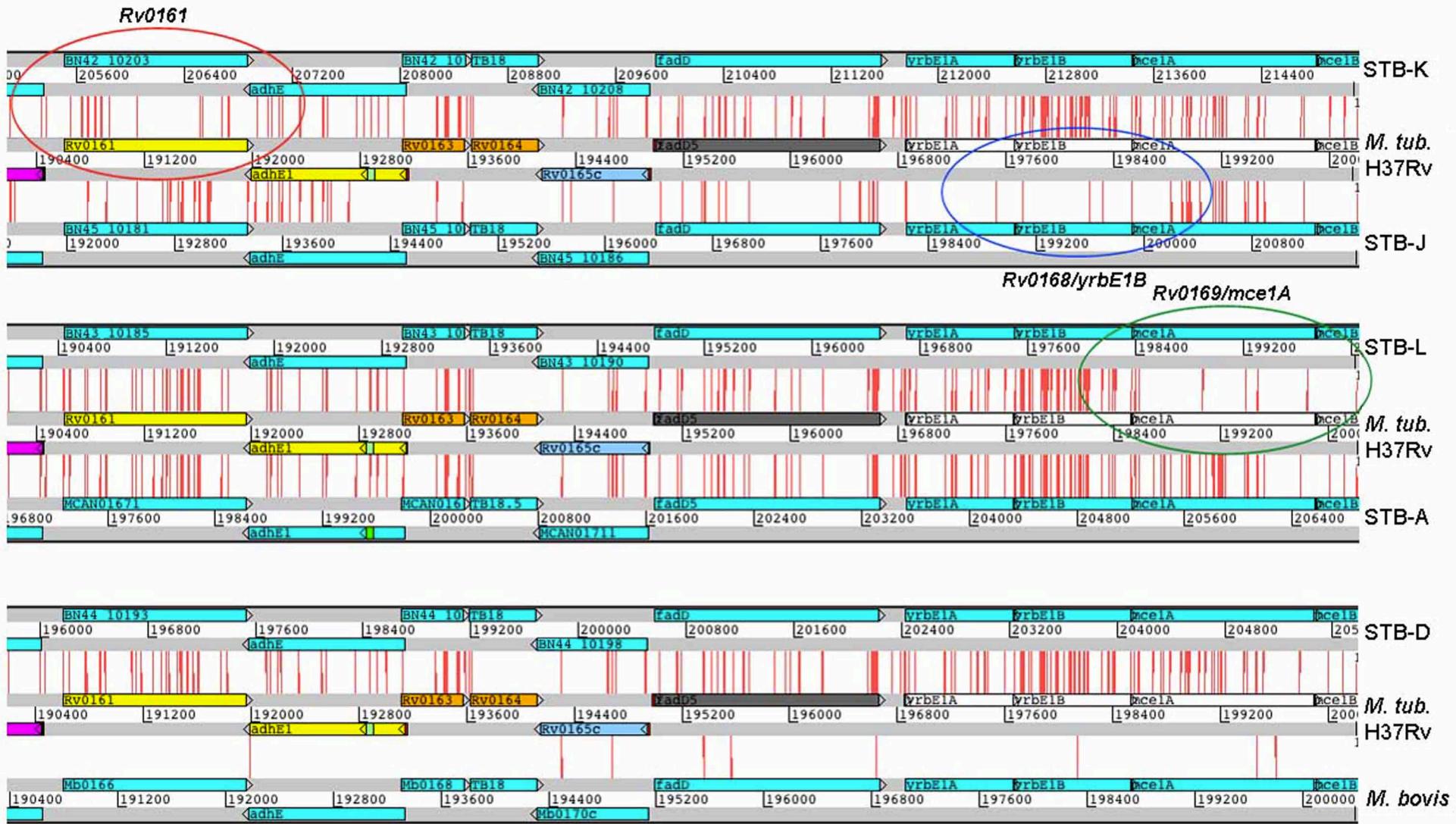
**Supplementary Fig. 4.** MTBC interrupted coding sequences intact in STB. Genomic comparison of STB strains with MTBC strains using ACT (upper panel) and MaGe (lower panel) for the display of interrupted coding sequences (ICDS). Example of gene *pks8*, whose open reading frame (ORF) was found intact in all STB strains and also in *M. marinum* and *M. avium*, while the orthologous gene in all MTBC members is interrupted due to a deletion of a thymine (T) residue. Incompletely colored boxes indicate interruption of Pks-encoding ORFs in MTBC strains in the lower panel. *M. tub.* = *M. tuberculosis*.



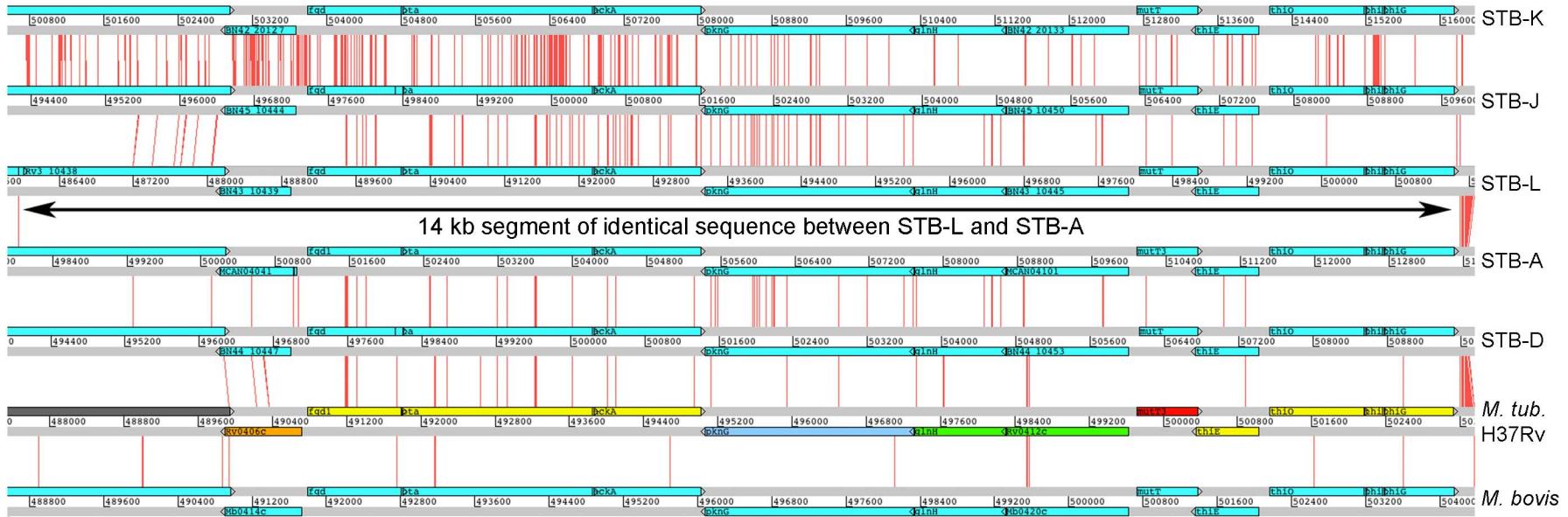
**Supplementary Fig. 5.** Ancestral gene organization of MTBC gene-fusions in STB. Example of a genetic locus where a presumably ancestral gene organization present both in *M. kansasii* and in most STB strains seems to have been rearranged to result in fusion gene *narX* in MTBC.



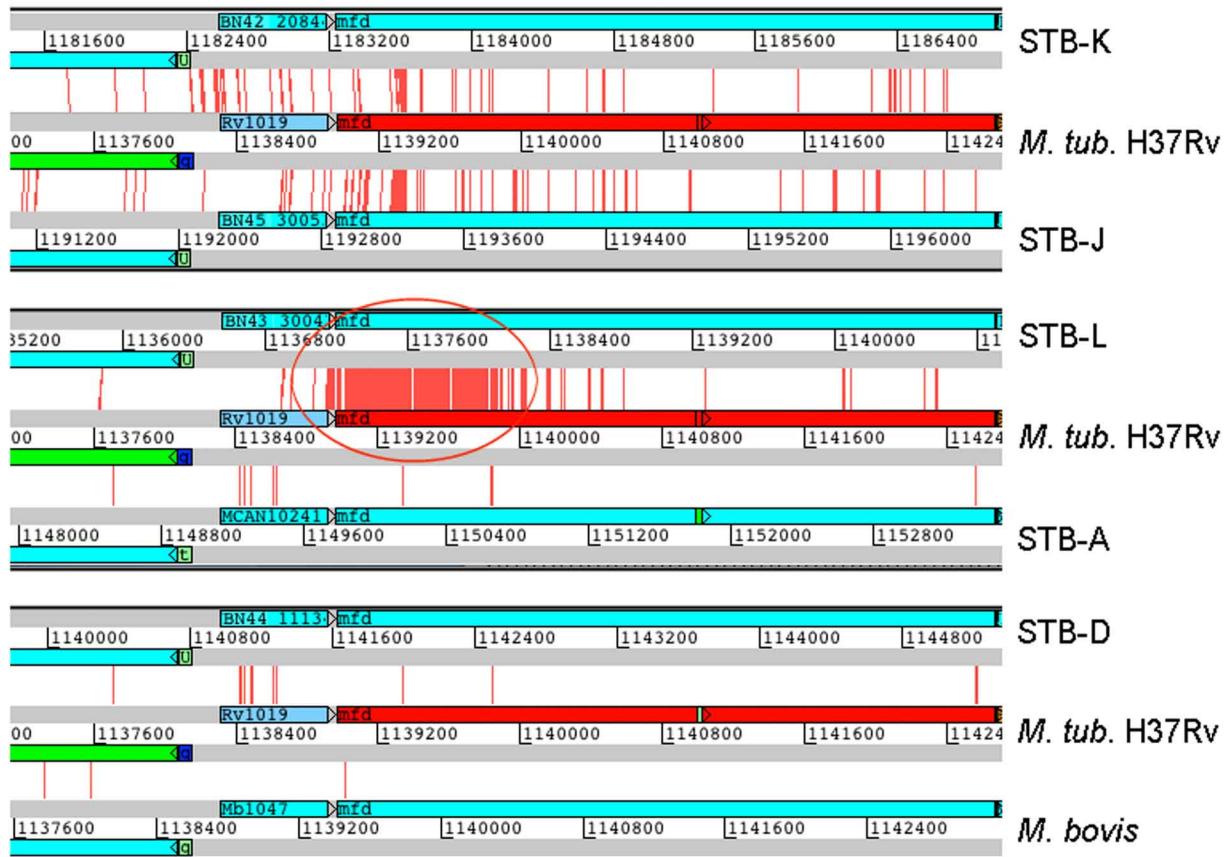
**Supplementary Fig. 6.** Network phylogeny of nine STB strains and *M. tuberculosis*. Network phylogeny inferred among the five STB isolates subjected to complete genome sequence analysis (STB-A, -D, -J, -K, and -L) and four additional STB isolates subjected to whole genome shotgun (WGS) sequencing (STB-E, -G, -H, and -I), in comparison to *M. tuberculosis* H37Rv. The phylogeny was established by NeighborNet analysis based on pairwise alignments of whole genome SNP data.



**Supplementary Fig. 7.** Inter-strain recombination regions in tubercle bacilli. **(a)** SNP distribution (represented as red lines) among STB and MTBC aligned genome segments showing likely inter-strain recombination regions with homoplasic similarities between strains that are marked with circles. Note that the presence of a few SNPs in these regions suggests ongoing diversification after the recombination events.



**Supplementary Fig. 7.** Inter-strain recombination regions in tubercle bacilli. **(b)** Representation of a genomic segment in the size of 14 kb that shows nucleotide sequence identity between STB-L and STB-A strains, representing a plausible recombination region involving the region encoding the serine/threonine protein kinase PknG.



>ref|ZP\_04748553.1| transcription-repair coupling factor Mfd (TrcF) [Mycobacterium kansasii ATCC 12478] Length=1238

Score = 498 bits (1282), Expect = 1e-163, Method: Compositional matrix adjust.  
Identities = 253/298 (85%), Positives = 275/298 (92%), Gaps = 0/298 (0%)

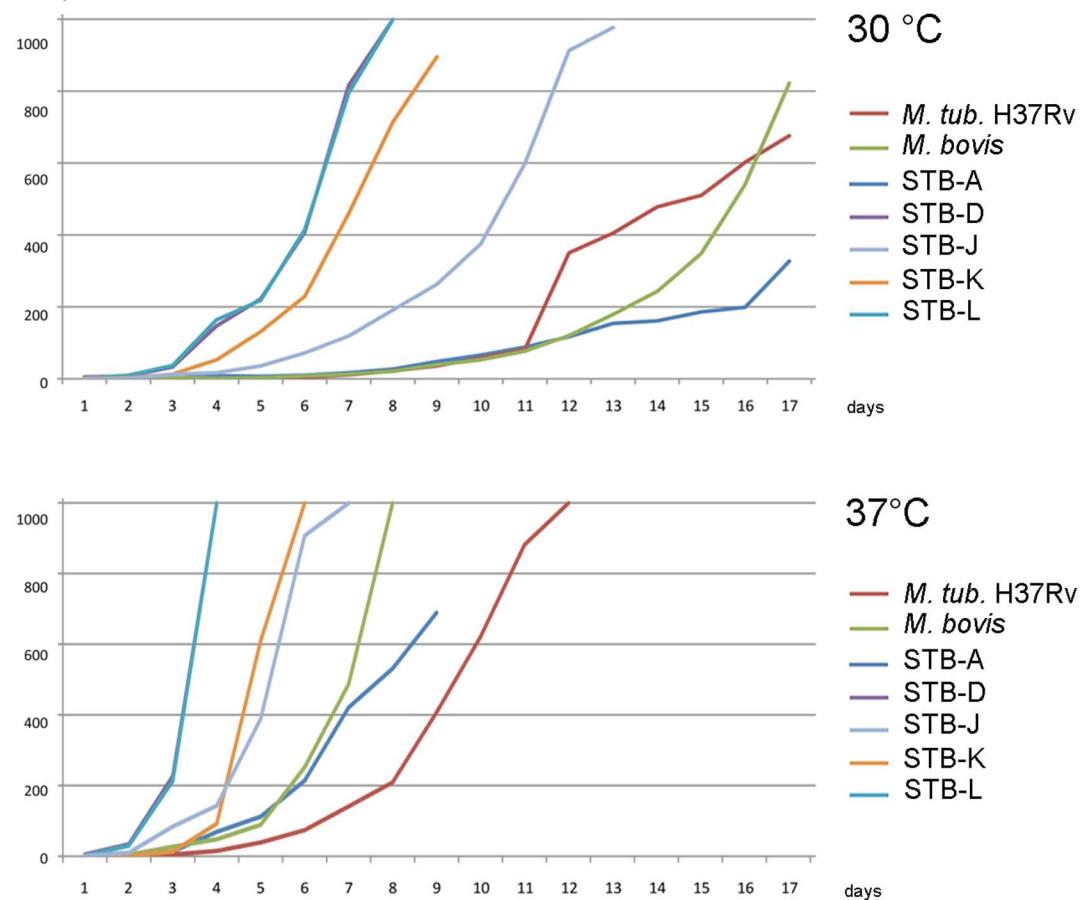
Query 1	MTAPGPGSPDTPPIAGLVALALRSPEFQQLIARSSDQPDELTLVGPAGARLFVASALAQRG	60
Sbjct 1	MTAPG SPDTPPIAGL LAL +P FQQLIAR+SD+PD+LTLVGP AARLFVASALA+ G	
Query 61	PLLVVTATGREADDLTAELRGVFGAAVAQFPSSWETLPHERLSPGVDTVGTRLMVLRLAH	120
Sbjct 61	PLLVVTATGREADDLTAEL+GV G AVA FPSWETLPHERLSPGVDTVG RL+VLLRLA+	
Query 121	PDDARLGPPLRVVVTAVRSLLQPMTPRLGQQEPITLSVGQEIGFEDVIARLVELAYTRVD	180
Sbjct 121	PDDARLGP SLRVVVTAVRSLLQPMTP+LG+QEP+ LSGVQE I+ VIARLVELAYTRVD	
Query 181	MVGRRGEFAVRGGILDIFAPTAEHPV RVEFWGDEITEMRMFSIADQRSIPGLDVDTLVAV	240
Sbjct 181	MVGRRGEFAVRGGILD+FAPTAEHPV RVEFWGDEITEMRMFS+ADQRSIP ++VDTLVAV	
Query 241	ACRELPLTDDVRARAQLAE QYPAAGDAITGSVTDM LAKLADGIPVGMEALFSVLAPG	298
Sbjct 241	ACREL LTDDVR RAA+LA Q+PAA +A+TGSV+DMLAKLA+GIPVGMEAL VL P	
	Sbjct 241 ACRELLS EDVRARAQLAARHPAAESTVTGSASDMLAKLAEGI A VDGMEA LPV LWDG	298

GENE ID: 886077 mfd | transcription-repair coupling factor  
[Mycobacterium tuberculosis H37Rv] (Over 10 PubMed links)

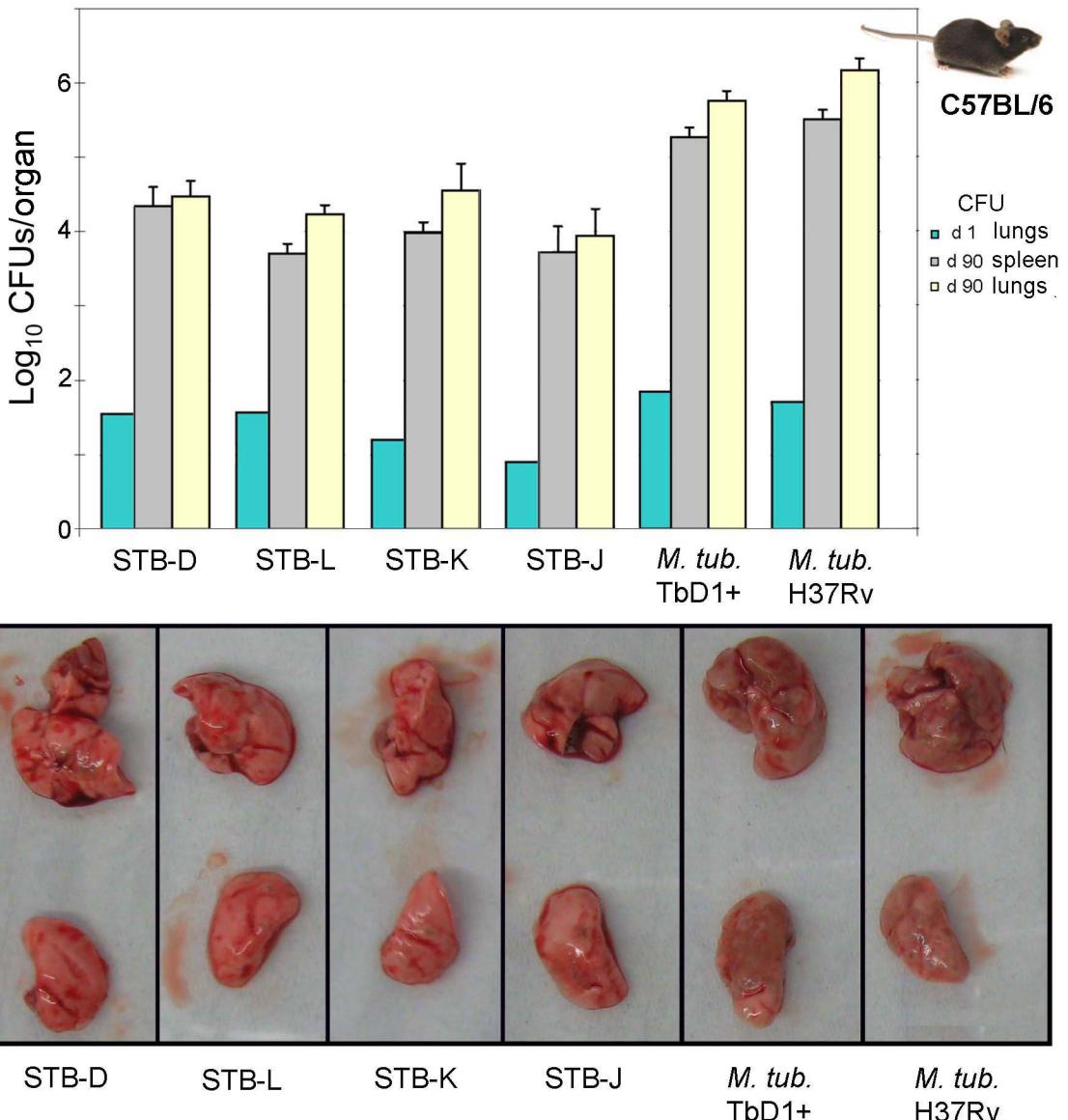
Score = 447 bits (1150), Expect = 3e-144, Method: Compositional matrix adjust.  
Identities = 236/300 (79%), Positives = 261/300 (87%), Gaps = 0/300 (0%)

Query 1	MTAPGPGSPDTPPIAGLVALALRSPEFQQLIARSSDQPDELTLVGPAGARLFVASALAQRG	60
Sbjct 1	MTAPG DTPPIAGL LAL +P FQQL+ R+ +PDETL+ PA ARL VASALA++G	
Query 61	PLLVVTATGREADDLTAELRGVFGAAVAQFPSSWETLPHERLSPGVDTVGTRLMVLRLAH	120
Sbjct 61	PLLVVTATGREADDL AELRGVFG AVA PSWETLPHERLSPGVDTVGTRLM LRLAH	
Query 121	PDDARLGPPLRVVVTAVRSLLQPMTPRLGQQEPITLSVGQEIGFEDVIARLVELAYTRVD	180
Sbjct 121	PDDAQLGPP LVVVTAVRSLLQPMTP+LG EP+TL+VG E F+ V+ARLVELAYTRVD	
Query 181	MVGRRGEFAVRGGILDIFAPTAEHPV RVEFWGDEITEMRMFSIADQRSIPGLDVDTLVAV	240
Sbjct 181	MVGRRGEFAVRGGILD+FAPTAEHPV RVEFWGDEITEMRMFS+ADQRSIP ++VDTLVAV	
Query 241	ACRELPLTDDVRARAQLAE QYPAAGDAITGSVTDM LAKLADGIPVGMEALFSVLAPG	300
Sbjct 241 ACREL L+DVRARAQLAQ ++PAA +TGS +DMLAKLA+GI VDGMEA+ VL G		
	Sbjct 241 ACRELLSEDVRARAQLAARHPAAESTVTGSASDMLAKLAEGI A VDGMEA LPV LWDG	300

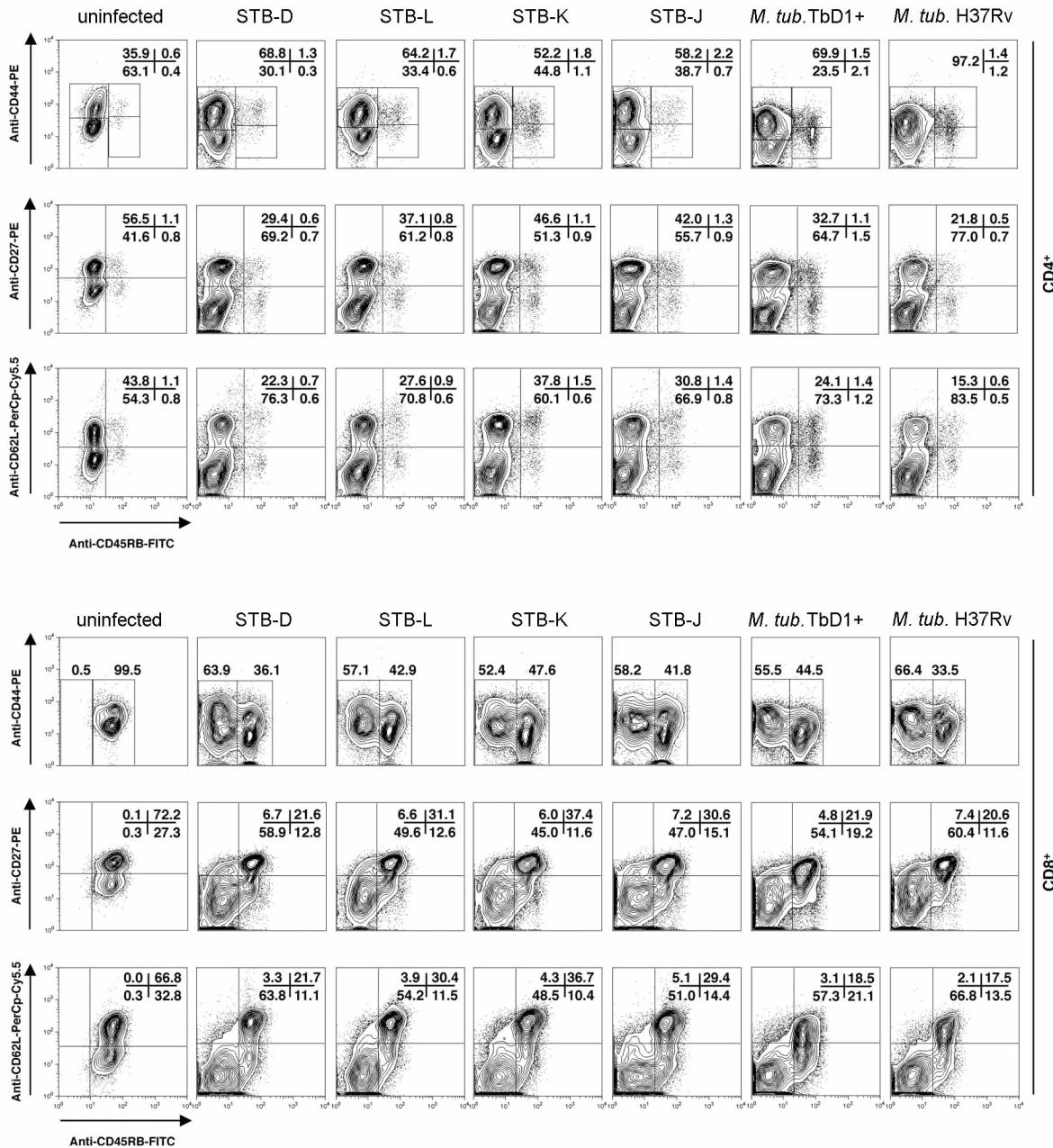
**Supplementary Fig. 7.** Inter-strain recombination regions in tubercle bacilli. (c) SNP distribution among STB and MTBC aligned genome segments, showing a likely exogenous importation/recombination region involving part of the *mfd* (*rv1020*) orthologue of STB-L. Sequence comparison of this region showed best hits with *M. kansasii* (85% amino acid sequence identity) that were higher than those with *M. tuberculosis* (79% identity).



**Supplementary Fig. 8.** *In vitro* growth characteristics of STB and MTBC members.  
 Density of bacterial cultures / growth curves of STB and MTBC members at 30 and 37°C  
 as measured by BACTEC 460 system (Beckton-Dickinson) in numerical growth units.



**Supplementary Fig. 9.** Virulence of STB and *M. tuberculosis* strains in C57BL/6 mice. Colony forming units (CFUs) recovered from lungs and spleens after aerosol infection 13 weeks after infection, compared to the inhaled dose present in the lungs at day 1. Solutions used for nebulization were concentrated as follows: STB-D ( $3.7 \times 10^6$ ), STB-L ( $3.6 \times 10^6$ ), STB-K ( $4.1 \times 10^6$ ), STB-J ( $1.1 \times 10^6$ ), *M. tuberculosis* TbD1+ ( $3.3 \times 10^6$ ) and *M. tuberculosis* H37Rv ( $3.6 \times 10^6$ ). The CFUs shown correspond to the average counts obtained for 4 mice per group. The lower panel shows that in accordance with CFU counts lungs from mice infected with *M. tuberculosis* (TbD1+ / H37Rv) show more severe lesions than lungs from mice infected with STB strains.



**Supplementary Fig. 10.** Immune responses induced by STB and *M. tuberculosis* strains. Profile of recruitment and activation of adaptive immune cells in the lung parenchyma of C57BL/6 mice infected with different STB or *M. tuberculosis* strains. Mice ( $n = 4/\text{group}$ ) were infected via aerosol with  $\sim 100 \text{ CFU}/\text{mouse}$  of strains STB-D, -L, -K, -J or *M. tuberculosis* (TbD1+ / H37Rv) or left untreated. The profiles of T-cell activation, i.e. CD45RB down regulation, CD44 modulation, CD27 down regulation, and of T-cell migration, i.e. CD62L down regulation of lung CD4+ (upper panel) or CD8+ (lower panel) T subsets were evaluated by cytofluorometric analyses at 13 weeks post-infection. Pools of lung cells from at least two representative experiments are shown.

**Supplementary Table 1. Bacterial strains used in this study**

NRL code	Secondary code	Classification	ST <sup>a</sup>	Site of isolation	Date of isolation	Country of isolation	Genome sequenced
19990160		STB-canetti	CD	pulmonary	1999	France	
19990161		STB-canetti	CD	lymph node	1999	Djibouti	
19990263		STB	F	pulmonary	1997	France	
19990264		STB	F	pulmonary	1998	France	
19990515		STB-canetti	CD	pulmonary	1999	Djibouti	
19990516	CIPT 140060008	STB-canetti	CD	pulmonary	1999	Djibouti	X
19990589		STB-canetti	CD	pulmonary	1999	Djibouti	
19990645		STB	L	lymph node	1997	Djibouti	
19990711		STB-canetti	B	lymph node	1999	Djibouti	
19990768	CIPT 140070017	STB	J	lymph node	1999	Djibouti	X
19991574		STB-canetti	CD	pulmonary	1999	France	
19991669		STB	N	pulmonary	1999	Djibouti	
19991704		STB-canetti	CD	pulmonary	1999	Djibouti	
19991705		STB	H	pulmonary	1999	Djibouti	
19991708		STB-canetti	CD	lymph node	1999	Djibouti	
19991709	CIPT 140070002	STB	E	lymph node	1999	Djibouti	*
20000239		STB	CD	bone	1999	France	
19970130		STB-canetti	CD	pulmonary	1997	France	
20000342		STB-canetti	CD		2000	Djibouti	
20000473	CIPT 140070007	STB	I	pulmonary	2000	Djibouti	*
20000586		STB-canetti	CD	pulmonary	2000	Djibouti	
20000587	CIPT 140070005	STB	G	lymph node	2000	Djibouti	*
20001049		STB-canetti	CD	pulmonary	2001	Djibouti	
20001155		STB-canetti	CD	pulmonary	2000	France	
20001245		STB-canetti	CD	pulmonary	2000	Djibouti	
20001246		STB-canetti	CD	pulmonary	2000	Djibouti	
20001247		STB-canetti	CD	peritoneal liq.	2000	Djibouti	
20001248		STB-canetti	CD	pulmonary	2000	Djibouti	
19980862		STB-canetti	CD	lymph node	1998	Djibouti	
20010188		STB-canetti	CD	lymph node	2001	Djibouti	
20010389		STB-canetti	CD	peritoneal liq.	2001	Djibouti	
20010390		STB-canetti	CD	lymph node	2001	Djibouti	
20010391		STB	F	pulmonary	2001	Djibouti	
20010933		STB-canetti	CD	lymph node	2001	Djibouti	
20020544		STB-canetti	CD	pulmonary	2002	Djibouti	
20020986		STB-canetti	CD	lymph node	2002	Djibouti	
20020987		STB-canetti	CD	bone	2002	Djibouti	
20020988		STB-canetti	CD	pulmonary	2002	Djibouti	
20020989		STB-canetti	CD	pulmonary	2002	Djibouti	
19980863	CIPT 140070013	STB	H	pulmonary	1998	Djibouti	*
20021261		STB-canetti	CD	lymph node	2002	Djibouti	
20030159		STB-canetti	CD	pulmonary	2002	France	
20030466		STB-canetti	CD	lymph node	2002	Djibouti	
20030467		STB-canetti	CD	blood	2003	Djibouti	
20030686		STB-canetti	CD	blood	2003	Djibouti	
20033147		STB-canetti	CD	lymph node	2003	France	
20040352		STB	M	lymph node	2004	Djibouti	
20041158	CIPT 140070010	STB	K	Skin	2004	Djibouti	X
20050462	CIPT 140070008	STB	L	lymph node	1997	Djibouti	X
20050642		STB-canetti	CD	pulmonary	2005	Djibouti	
140010059	CIPT140060001	STB-canetti	A	pulmonary	1969	France	X
140010060		STB-canetti	A	pulmonary	1969	France	
140010061		STB-canetti	A	pulmonary	1970	Papeete	
19981514		STB-canetti	CD	lymph node	1998	Djibouti	
19990121		STB-canetti	CD	lymph node	1993	Switzerland	
140030001		<i>M. africanum</i>	MTBC				
	AF2122/97	<i>M. bovis</i>	MTBC	cow lymph node	1997	United Kingdom	
140080012		<i>M. caprae</i>	MTBC				
140050001	OV254	<i>M. microti</i>	MTBC	vole	1930	United Kingdom	
140090001		<i>M. pinnipedii</i>	MTBC	seal			
	210	<i>M. tuberculosis</i>	MTBC	pulmonary		United States	
CDC1551		<i>M. tuberculosis</i>	MTBC	pulmonary		United States	
H37Rv		<i>M. tuberculosis</i>	MTBC	pulmonary	1934		
Uganda		<i>M. tuberculosis</i>	MTBC	pulmonary		Uganda	
20030423	TbD1+	<i>M. tuberculosis</i>	MTBC	pulmonary	2003		

<sup>a</sup>Sequence type as determined by 12-housekeeping gene-based multilocus sequence typing (see also Supplementary Table 2). NRL, French National Reference Laboratory. All smooth tubercle bacilli (STB) are human clinical isolates. Note that the spectrum and distribution of tuberculosis forms (i.e. pulmonary vs. extrapulmonary) for STB isolates were similar to those observed among tuberculosis patients in this geographic region infected by classical MTBC strains<sup>1</sup>.

STB-canetti = smooth tubercle bacilli (STB) belonging to the *M. canetti*-like strain cluster; CIPT = Collection Institut Pasteur, Tuberculose;

X Whole genome sequencing performed by using Sanger, 454 and Illumina technologies, finishing and expert annotation

\* Genome sequencing performed by Illumina HiSeq (Whole Genome Shotgun) WGS

<sup>1</sup>Koeck, J.L. et al. Epidemiology of resistance to antituberculosis drugs in *Mycobacterium tuberculosis* complex strains isolated from adenopathies in Djibouti. Prospective study carried out in 1999. Med Trop 62, 70-72 (2002).

**Supplementary Table 2. PCR primer sequences used for multilocus sequence typing (MLST)**

Gene	H37Rv-No.	Forward and Reverse primers (5' to 3')	Amplicon size (bp)	Gene positions retained for MLST	Edited sequence size (bp)
<i>adk</i>	<i>rv0733</i>	GATCTCCACCGGGCGAACTCTT TACTTTCCCAGAGCCCGAAC	462	196-471	276
<i>pgi</i>	<i>rv0946c</i>	CGGCAGATCTCATCGACTA CGACAAGTCATTGGAATACGG	892	250-814	565
<i>pncA</i>	<i>rv2043c</i>	GATCATCGTCGACGTGCAGAA CAGGAGCTGAAACCAACTCG	549	100-475	376
<i>glyS</i>	<i>rv2357c</i>	AGAGAACATCAAGGCCAGTG CTTATCCATCCCACCCCTTGGT	928	304-855	552
<i>efp</i>	<i>rv2534c</i>	CCACTGCTGACTTCAAGAACG GCAGAATCCACCTTAGTTGTC	522	91-402	312
<i>gltS</i>	<i>rv2992c</i>	CCCAAGCTGGGTTACGACAAT CCAGTCCGTACACTTGTCA	888	508-1096	589
<i>gyrA</i>	<i>rv0006</i>	GTTCGTGTGTTGCGTCAAGT CAGCTGGGTGTGCTTGTAAA	1013	220-960	741
<i>katG</i>	<i>rv1908c</i>	CTACCAGCACCGTCATCTCA AGGTCGTATGGACGAACACC	913	1300-1851	552
<i>gyrB</i>	<i>rv0005</i>	TCGGACGCGTATGCGATATC ACATACAGTCGGACTTGC	1020	502-1437	936
<i>rpoB</i>	<i>rv0667</i>	GGATGTTGATCAGGGTCTGC TCAAGGAGAACGCGTACGA	340	898-1206	309
<i>hsp65</i>	<i>rv0440</i>	ACCAACGATGGTGTGTCAT CTTGTGAAACCGCATACCC	421	193-564	372
<i>sodA</i>	<i>rv3846</i>	AGCTTCACCAAGCAAGCACCA GCCAGTTACGACGTTCCAAA	46	205-516	312

All PCR fragments were sequenced on both strands using Sanger sequencing and an ABI 3700 DNA Analyzer.

**Supplementary Table 3.** Alignment of orthologues among MTBC and STB genomes  
(Please refer to the separate online file in .xls format)

**Supplementary Table 4. Genes present in MTBC genomes and fully or partially absent from STB genomes.**

Gene-locus label in <i>M. tuberculosis</i> H37Rv	Product	BLASTP results (excluding MTBC)
<i>rv0394c</i>	Possible Secreted Protein	Best Hit = 74/212 (35%) <i>Gordonia otitidis</i>
<i>rv0395</i>	Hypothetical Protein	Best Hit = 85/119 (71%) <i>Mycobacterium colombiense</i>
<i>rv0396</i>	Hypothetical Protein	Best Hit = 90/130 (69%) <i>Mycobacterium colombiense</i>
<i>rv0602c</i>	Probable Two Component DNA Binding Transcriptional Regulatory Protein TcrA	Best Hit = 142/225 (63%) <i>Patulibacter sp.</i>
<i>rv0603</i>	Possible Exported Protein	Best Hit = 49/93 (53%) <i>Arthrobacter sp. Chr15</i>
<i>rv0604</i>	Probable Conserved Lipoprotein LpqO	Best Hit = 171/308 (56%) <i>Arthrobacter sp. Chr15</i>
<i>rv0746</i>	PE_PGRS Family Protein	ND
<i>rv0755a</i>	Putative Transposase (Fragment)	ND
<i>rv0872c</i>	PE_PGRS Family Protein	ND
<i>rv0874c</i>	Conserved Hypothetical Protein	Best Hit = 310/383 (81%) <i>Mycobacterium kansasi</i>
<i>rv1046c</i>	Hypothetical Protein	Best Hit = 26/95 (27%) <i>Cupriavidus basilensis</i>
<i>rv1573-rv1586c</i> (14 genes)	Probable PhiRv1 Prophage-like Proteins	ND
<i>rv1818c</i>	PE_PGRS33, PE_PGRS family Protein	ND
<i>rv1989c</i>	Hypothetical Protein	Best Hit = 158/184 (86%) <i>Mycobacterium gilvum</i> plasmid pMFLV01; <i>Mycobacterium</i> sp. KMS plasmid pMKMS01
<i>rv1990c</i>	Probable Transcriptional Regulatory Protein	Best Hit = 105/113 (93%) <i>Mycobacterium</i> sp. JLS
		Best Hit = 103/113 (91%) <i>Mycobacterium gilvum</i> plasmid pMFLV01, <i>Mycobacterium</i> sp. KMS plasmid pMKMS01
<i>rv1990a</i>	Possible Dehydrogenase (Fragment)	ND
<i>rv2023c</i>	Hypothetical Protein	Best Hit = 25/84 (30%) <i>Candidatus Kuenenia stuttgartiensis</i>
<i>rv2023a</i>	Hypothetical Protein	ND
<i>rv2645</i>	Hypothetical Protein	Best Hit = 34/103 (33%) <i>Robiginitalea biformata</i>
<i>rv2646</i>	Probable Phage Integrase	Best Hit = 166/255(65%) <i>Mycobacterium kansasi</i>
<i>rv2647</i>	Hypothetical Protein	Best Hit = 38/117 (32%) <i>Nocardioidaceae bacterium</i>
<i>rv2650c-rv2659c</i> (10 genes)	Probable PhiRv2 Prophage-like Proteins	ND
<i>rv3190c</i>	Hypothetical Protein	Best Hit = 195/420 (46%) <i>Pseudonocardia dioxanivorans</i>
<i>rv3386</i>	Possible Transposase (/S1560)	ND
<i>rv3387</i>	Possible Transposase (/S1560)	ND
<i>rv3388</i>	PE_PGRS Family Protein	ND
<i>rv3513c</i>	Probable Fatty-Acid-CoA Ligase FadD18 (Fragment) (Fatty-Acid-CoA Synthetase)	ND
<i>rv3590c</i>	PE-PGRS Family Protein	ND
<i>rv3902c</i>	Hypothetical Protein	Best Hit = 67/175 (38%) <i>Mycobacterium avium</i> subsp. <i>paratuberculosis</i>

ND = not determined on genes potentially containing repetitions.

**Supplementary Table 5. Gene scars detected in MTBC and STB strains**

Scar n°	Scar specificity	H37Rv ICDS		Orthologous genes in					Putative function	Functional classification	Causative mutation
		STB-A	STB-D	STB-J	STB-K	STB-L	<i>M. marinum</i>				
1	MTBC	<i>rv0890c-rv0891c</i>	<i>MCAN_08911 BN44_10973 BN45_20182 BN42_20676 BN43_20347</i>				Region absent but complete paralogs in e.g. <i>M. leprae</i>	Transcriptional regulator	Regulation	Frameshift, 1-nt insertion in MTBC	
2	MTBC	<i>rv1203c-rv1204c</i>	<i>MCAN_12181 BN44_11344 BN45_30265 BN42_21068 BN43_30272</i>				<i>MMAR4234</i>	Unknown	Conserved hypothetical	Non sense mutation in MTBC	
3	MTBC	<i>rv1662-rv1663</i>	<i>MCAN_16711 BN44_20224 BN45_40135 BN42_21590 BN43_30781</i>				<i>MMAR2472</i>	Polyketide synthase Pks8/17	Lipid metabolism	Frameshift, 1-nt deletion in MTBC	
4	MTBC	<i>rv3741c-rv3742c</i>	<i>MCAN_37621 BN44_120145</i>	Region absent	<i>BN42_90261 BN43_90250</i>		<i>MMAR5279</i>	Unknown	Conserved hypothetical	Frameshift, 1-nt deletion in MTBC	
5	Tubercle bacilli (MTBC + STB)	<i>rv1104-na-rv1105</i>	<i>MCAN_11141-11151 BN44_11228-11229-11230 BN45_30146-30147-30148-30149 BN42_20951-20952-20953 BN43_30160-30161-30162</i>				<i>MMAR4363</i>	Para-nitrobenzyl esterase	Intermediary metabolism	Frameshifts in MTBC and STB	
6	Tubercle bacilli (MTBC + STB)	<i>rv2321c-rv2322c</i>	<i>MCAN_23521-23531 BN44_50281-50282 BN45_50678-50679 BN42_40246-40247 BN43_31580-31581</i>				<i>MMAR3622</i>	Ornithine aminotransferase RocD	Intermediary metabolism	Frameshift, 1-nt deletion in MTBC and STB	
7 to 81	Mycobacterium or beyond	various	various	various	various	various	various	various	various	various	various

ICDS, interrupted coding sequence<sup>1</sup>; na, not annotated ICDS; nt, nucleotide. Green cells show ICDS in MTBC that correspond to intact coding sequence (CDS) both in smooth tubercle bacilli (STB), *M. marinum* and other mycobacteria, indicating scars that likely occurred in the most recent common ancestor (MRCA) of the MTBC after divergence from STB-like progenitors. Blue cells show ICDS shared by MTBC and STB and intact CDS in *M. marinum* and other mycobacteria, indicating scars that likely occurred in the most recent common ancestor (MRCA) of all tubercle bacilli (MTBC and STB) after divergence from other mycobacterial species. Purple cells show ICDS found both in MTBC, STB and more distantly related (myco)bacteria, suggesting evolutionary more ancient scars.

<sup>1</sup>Deshayes, C. et al. Detecting the molecular scars of evolution in the *Mycobacterium tuberculosis* complex by analyzing interrupted coding sequences. BMC Evol Biol 8, 78 (2008).

**Supplementary Table 6. Genomic regions showing homoplastic similarities between strains.**

The first column depicts strains whose sequences display homoplastic similarities in certain genomic regions. In the second column the genes corresponding to these homoplastic regions are listed. Locus tags were collected from the annotated genome sequence of the underlined strain (when available, the gene name is also provided in parentheses). Contiguous gene loci are shown on the same line.

Homoplastic similarity	Locus tags of affected genes
<u>M. tub.</u> H37Rv + <u>M. bovis</u> + STB-J	<i>rv0151c-rv0152c,</i> <i>rv0154c (fadE2),</i> <i>rv0168 (yrbE1B),</i> <i>rv0198c,</i> <i>rv0355c,</i> <i>rv0356c-rv0357c (purA),</i> <i>rv0755c,</i> <i>rv1889c,</i> <i>rv1936-rv1937,</i> <i>rv2064 (cobG),</i> <i>rv2515c,</i> <i>rv2793c (truB)-rv2797c,</i> <i>rv3025c (iscS)-rv3026c,</i> <i>rv3596c (clpC1)-rv3598c (lysS),</i> <i>rv3600c,</i> <i>rv3896c-rv3897c,</i>
<u>M. tub.</u> H37Rv + <u>M. bovis</u> + STB-K	<i>rv0161,</i> <i>rv0255c (cobQ1),</i>
<u>M. tub.</u> H37Rv + <u>M. bovis</u> + STB-L	<i>rv0169 (mce1a),</i> <i>rv0425c (ctpH)-rv0426c,</i> <i>rv0959,</i> <i>rv2520c,</i> <i>rv3034c,</i>
STB-A + STB-D + STB-J	<i>MCAN_12161,</i> <i>MCAN_38011 (rfbE)-MCAN_38021,</i>
STB-A + STB-D + STB-K	<i>MCAN_00061,</i> <i>MCAN_00941-MCAN_00961 (ctpA),</i> <i>MCAN_00991,</i> <i>MCAN_01801 (mce1F),</i> <i>MCAN_01861,</i> <i>MCAN_02031,</i> <i>MCAN_04211 (thiC),</i> <i>MCAN_04231 (ctpH)-MCAN_04241,</i> <i>MCAN_08361,</i> <i>MCAN_13911,</i> <i>MCAN_16101 (hisH),</i> <i>MCAN_16121 (impA),</i> <i>MCAN_21221,</i> <i>MCAN_21271 (helZ)-MCAN_21281,</i> <i>MCAN_21351,</i> <i>MCAN_21401 (ippK),</i> <i>MCAN_24481-MCAN_24491,</i> <i>MCAN_30521 MCAN_30531 (fixB),</i> <i>MCAN_30561,</i> <i>MCAN_32401-MCAN_32421-MCAN_32441,</i> <i>MCAN_32461 (aroA)-MCAN_32471</i>

<u>STB-A + STB-D + STB-L</u>	<i>MCAN_02841,</i> <i>MCAN_03061,</i> <i>MCAN_06101,</i> <i>MCAN_20871 (cobG)-MCAN_20891(cobI),</i>
<u>STB-A + STB-J</u>	<i>MCAN_05651 (grcC1)-MCAN_05661 (htpX),</i> <i>MCAN_06471,</i> <i>MCAN_07621 (phoR),</i> <i>MCAN_07741,</i> <i>MCAN_07901-MCAN_07921 (purQ),</i> <i>MCAN_07951-MCAN_07961,</i> <i>MCAN_08001-MCAN_08011 (pepC),</i> <i>MCAN_08041 (purl),</i> <i>MCAN_15921 (bioF1),</i> <i>MCAN_17221,</i> <i>MCAN_19511 (fadE17)-MCAN_19521 (echA13),</i> <i>MCAN_20171 (ctpF)-MCAN_20181,</i> <i>MCAN_20491-MCAN_20521 (pfkB),</i> <i>MCAN_20561,</i> <i>MCAN_37441,</i> <i>MCAN_37871-MCAN_37881,</i> <i>MCAN_37961-MCAN_37971,</i> <i>MCAN_38411,</i>
<u>STB-A + STB-K</u>	<i>MCAN_01021 (fadD10),</i> <i>MCAN_05691-MCAN_05701</i> <i>MCAN_05771 MCAN_05781</i> <i>MCAN_06731 (echA4)-MCAN_06751 (echA5),</i> <i>MCAN_14311 (ribA2)- MCAN_14351,</i> <i>MCAN_14531 (pgk)-MCAN_14551,</i> <i>MCAN_16371 (polA)</i> <i>MCAN_19341,</i> <i>MCAN_20711 (pkS12),</i> <i>MCAN_33171,</i> <i>MCAN_34011 (amiD)-MCAN_34021,</i> <i>MCAN_34161 (iunH)-MCAN_34171,</i> <i>MCAN_34851 (mhpE)-MCAN_34861 (ilvB2)</i> <i>MCAN_35151 (yrbE4A)-MCAN_35161,</i> <i>MCAN_35181 (fadE26)-MCAN_35191 (fadE27),</i> <i>MCAN_35381,</i> <i>MCAN_35801, MCAN_35811 (bphC),</i> <i>MCAN_37291,</i>
<u>STB-A + STB-L</u>	<i>MCAN_04081 (pknG) region,</i> <i>MCAN_06621 (atsD),</i> <i>MCAN_10671,</i> <i>MCAN_20741 (ppm1),</i> <i>MCAN_32961,</i>
<u>STB-K + STB-D</u>	<i>BN42_21457,</i> <i>BN42_21589,</i> <i>BN42_30192 (cinA),</i> <i>BN42_50163,</i> <i>BN42_90228 (leuA),</i> <i>BN42_90444,</i>

STB-K + STB-L

*BN42\_21026 (narH)-BN42\_21027 (narJ),  
BN42\_21058-BN42\_21059 (fadD),  
BN42\_21302,  
BN42\_21391 (ctpD),  
BN42\_21559,  
BN42\_21578 (pheT),  
BN42\_30209,  
BN42\_30215 (lipD),  
BN42\_40041,  
BN42\_40313 (mbtE),  
BN42\_40411 (lipP),  
BN42\_40821 (ugpA)-BN42\_40824 (rbfA),  
BN42\_40930 (ppsB),  
BN42\_40938 (mas),  
BN42\_41339,  
BN42\_50034,  
BN42\_90429 (eccC).*

---

*M. tub.* = *M. tuberculosis*, STB = smooth tubercle bacilli.

**Supplementary Table 7. Histopathological analyses of BALB/c mice infected by *Mycobacterium tuberculosis* or STB strains.** Six groups, of three 7-week old female BALB/c mice each, were infected intranasally by either of the strains indicated below at a dose of  $10^3$  colony forming units. After 128 days post-infection, mice were sacrificed and the lungs of each mouse were fixed in 4 % paraformaldehyde and embedded in paraffin *in toto*. Histological sections of 5  $\mu\text{m}$  including the totality of the pulmonary lobes were stained using hematoxylin-eosin or Ziehl-Neelsen staining. The totality of the sections were examined and graded for the severity of lesions as indicated below the table.

Inflammation <sup>b</sup>	Mice infected by																																			
	<i>M. tuberculosis</i> H37Rv						STB-A						STB-D						STB-L						STB-K						STB-J					
	Mouse 1	Mouse 2	Mouse 3	Incidence	Average <sup>d</sup>	Std deviation	Mouse 1	Mouse 2	Mouse 3	Incidence	Average <sup>d</sup>	Std deviation	Mouse 1	Mouse 2	Mouse 3	Incidence	Average <sup>d</sup>	Std deviation	Mouse 1	Mouse 2	Mouse 3	Incidence	Average <sup>d</sup>	Std deviation	Mouse 1	Mouse 2	Mouse 3	Incidence	Average <sup>d</sup>	Std deviation	Mouse 1	Mouse 2	Mouse 3	Incidence	Average <sup>d</sup>	Std deviation
Extension of lesions <sup>a</sup>	3	3	3	3/3	3,0	0,0	1	1	1	3/3	1,0	0,0	2	3	2	3/3	2,3	0,6	1	1	1	3/3	1,0	0,0	2	1	1	3/3	1,3	0,6	1	1	2	3/3	1,3	0,6
Macrophages	3	3	3	3/3	3,0	0,0	1	1	1	3/3	1,0	0,0	2	3	3	3/3	2,7	0,6	1	1	1	3/3	1,0	0,0	2	1	1	3/3	1,3	0,6	2	2	3	3/3	2,3	0,6
Epitheloid cells	3	3	3	3/3	3,0	0,0	1	1	1	3/3	1,0	0,0	2	3	3	3/3	2,7	0,6	1	1	1	3/3	1,0	0,0	2	1	1	3/3	1,3	0,6	2	2	3	3/3	2,3	0,6
Foamy macrophages	2	2	2	3/3	2,0	0,0	1	1	1	3/3	1,0	0,0	2	2	2	3/3	2,0	0,0	1	1	1	3/3	1,0	0,0	1	1	1	3/3	1,0	0,0	1	1	1	3/3	1,3	0,6
Lymphocytes/plasmacytoid dendritic cells	3	3	3	3/3	3,0	0,0	1	2	2	3/3	1,7	0,6	2	3	3	3/3	2,7	0,6	2	2	2	3/3	2,0	0,0	2	2	2	3/3	2,0	0,0	2	2	3	3/3	2,3	0,6
Neutrophiles	2	2	2	3/3	2,0	0,0	1	1	#	2/3	1,0	0,0	1	#	1	2/3	1,0	0,0	1	1	1	3/3	1,0	0,0	1	1	1	3/3	1,0	0,0	1	1	#	2/3	1,0	0,0
Total inflammation <sup>e</sup>	16	16	16	3/3	16,0	0,0	6	7	6	3/3	6,3	0,6	11	14	14	3/3	13,0	1,7	6	7	7	3/3	6,7	0,6	1	7	7	3/3	8,0	1,7	9	9	13	3/3	10,3	2,3
Necrosis <sup>b</sup>	3	#	#	1/3	3,0	NA	#	#	#	0	NA	NA	#	#	#	0	NA	NA	#	#	#	0	NA	NA	#	#	#	0	NA	NA						
Acid fast bacilli <sup>b,c</sup>	2	2	1	2/3	1,7	0,6	0	0	0	0	NA	NA	0	0	0	0	NA	NA	0	0	0	0	NA	NA	0	0	0	0	NA	NA						

<sup>a</sup> scored according to the surface of the lung parenchyma observed at low magnification: 1 = 1-25%, 2 = 26-50%, 3 = 51-75%, 4 = 76-100%

<sup>b</sup> scored as follows: 1 = minimum to 5 = severe

<sup>c</sup> occurrence as detected after Ziehl-Neelsen staining

<sup>d</sup> averages were calculated only based on values from mice where lesions were observed

<sup>e</sup> Sum of all inflammation parameter scores

NA, not applicable; #, absent lesion